iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82787

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

MedGem AI: An Open-Source Multi-Modal Platform for Medical Image Analysis

Dr. U. M. Patil[1], Nomesh R. Kirange[2], Mansi S. Bendale[3], Tejas R. Jadhav[4], Krushna D. Patil[5]
[1] *Professor,* [2, 3, 4, 5]*Student, Department Of Computer Science and Engineering (Data Science), R. C. Patel Institute Of Technology, Shirpur, Maharashtra, India.*

ABSTRACT: ***Current healthcare systems around the world experience a tremendous bottleneck since more than 60% of the global population lacks access to prompt interpretations by specialist medical imaging experts. In order to solve this urgent problem, MedGem AI presents a locally installed, HIPAA-compliant platform that enables physicians to get intelligence for decision-making purposes. Using Google's sophisticated open-weighted medical AI models, MedGem AI aims to deliver expert insights in the most resource-constrained clinical settings.***

The core of the technology relies on a hierarchical architecture implemented under the Health AI Developer Foundations (HAI-DEF). Within this ecosystem, MedGemma 4B model takes the center stage and is responsible for interpreting imagery, dealing with visual questions, and automatically creating structured clinical notes. Other models used to accomplish different tasks include CXR Foundation (anomalies identification in chest X-ray images), Derm Foundation (diagnostic interpretation of skin lesions), and Path Foundation (classification of tissues in digital histopathology). To make this system work with regular consumer-grade hardware having at least 8GB of VRAM, NF4 4-bit quantization was used when developing each network.

Zero-Footprint Privacy Architecture represents the basis of the solution in terms of data sovereignty and security. To ensure that patient PHI does not go to the cloud, MedGem AI runs all computational workflows on-site. In addition, the developed Privacy Guard automatically removes metadata and PHI from DICOMs to keep this process HIPAA-compliant. To gain physicians' trust and increase robustness of the solution, MedGem AI implements Explainable AI capabilities by showing attention maps (where the model focuses) and uses Ensemble Fusion Engine for combining model predictions by means of weighted voting.

KEY WORDS: Multimodal AI, MedGemma, Clinical Decision Support (CDS), Privacy-Preserving Machine Learning, Medical Imaging, 4-bit Quantization, On-premise Deployment, Explainable AI (XAI),Chest Radiology, Digital Pathology

I. INTRODUCTION

As we saw in class, the implementation of AI algorithms in clinical workflows is triggered by the urgent need arising from a significant gap around the globe – 60% of the world's population lacks timely and accurate interpretations by specialists of medical imaging diagnostics. The problem comes from a shortage of pathologists and radiologists worldwide. The inability of patients in rural or poor areas to diagnose bone fractures or malignant tumors immediately contributes to bad health outcomes and increased cost of treatment.

As a result, in the beginning, medical AI technology tried to automate and optimize screening processes by applying deep learning CNNs. Models, including ResNet50 and DenseNet169, were used to detect anomalies in the database, and some managed to achieve similar levels of accuracy to human experts in certain categories like in MURA dataset.

However, the problem is that current CNNs work within the discriminative framework of operation. They can give the diagnostic probability or binary answer, while it's not possible to explain the reasoning behind their predictions or include any additional information about the patient's condition. It means that there are certain limitations to apply the current technology in practice because clinicians don't have an opportunity to identify what type of fracture the model found and what characteristics caused that.

Moreover, object detection models like YOLO (You Only Look Once) algorithm, despite the possibility to use the method in real-time, usually ignore micro-fractures, which might occur in low-contrast images for pediatric or geriatric cases. Therefore, the approach is still limited to treating images as mathematical tasks only and doesn't take into account patients' medical history.

The current trend in medical AI is shifting from classification models to Multimodal Large Language Models (MLLMs), where MedGemma, developed by Google, with two versions in 4B and 27B parameters, is leading the way. Unlike conventional classifiers that merely classify images, MedGemma is a multimodal assistant that can conduct complex reasoning and make contextual interpretations. By using a special MedSigLIP vision encoder along with a high-capacity language model, it is able to deal with Visual Question Answering (VQA) tasks without any difficulty.

Thus, doctors are provided with an opportunity to hold conversations with the system regarding various issues related to chest X-ray or pathology slides. Most importantly, since it is capable of creating natural text, it can formulate comprehensive clinical reports based on findings, impressions, and suggestions.

One of the reasons for which MedGemma is better than competing models in terms of performance is the method of its training. Typically, conventional computer vision models, including ResNet, are initially trained on generalized consumer datasets (such as the ImageNet database comprising thousands of common object pictures) and later fine-tuned on medical data. However, MedGemma is initially grounded in medicine, meaning that it is trained on more than 10 million PubMed abstracts, 2 million medical case notes, and millions of medical images. Consequently, the model receives advanced expertise in thearea, which makes it significantly outperform narrower models when dealing with out-of-distribution cases.

II. SIGNIFICANCE OF THE SYSTEM

The importance of this model is that it alone has the capability of democratizing global specialized care, ensuring absolute patient confidentiality, and avoiding prohibitively high costs of setting up infrastructure. With 4-bit quantization, the platform manages to condense complex medical decision-making into an efficient 4B-parameter model that runs seamlessly on inexpensive, off-the-shelf consumer computers instead of pricey server farms. This makes it possible for low-budget, rural healthcare facilities—which have more than 60% of the population underserved in diagnostic interpretation—to provide AI-assisted diagnostics locally, keeping patient information secure and reducing diagnosis times significantly.

III. LITERATURE SURVEY

Recent advancements in computational medicine have shown that domain-adapted foundation networks work better than generalized large language models for specialized clinical applications. Leading this change is Google's MedGemma framework, an open-weight multimodal large language model available in 4B and 27B parameter versions. This model continues the trend of instruction-tuning on large clinical note collections and medical text-image datasets (Buskila, 2026; Elden, 2025).

Unlike simple fine-tuning, MedGemma is built upon a solid medical foundation. It integrates a vision encoder tailored for the domain with a strong language backbone. This combination outperforms task-specific models as well as those handling different data types on tests like the MedQA-USMLE examinations (Buskila, 2026; Zhu et al., 2025). The model can generate free-form text and perform interactive Visual Question Answering (VQA). This allows for the automated creation of well-structured clinical documents and assessments of images (Shih, 2026; Zhu et al., 2025).

Importantly, for decentralized, privacy-focused projects like MedGem AI, community builds optimized through 4-bit quantization show significant improvements in diagnostic performance compared to standard general models. They work efficiently on consumer-grade hardware with as little as 8GB of VRAM (Buskila, 2026). This combination of computing efficiency and cross-disciplinary reasoning makes MedGemma a key player in expanding expert-level clinical decision support to resource-limited settings while maintaining strict data control.

IV. METHODOLOGY

The design and implementation of the MedGEM AI platform took place through a clear, four-stage engineering process. This approach gradually developed the system from basic, task-specific computer vision models into a complex, privacy-focused multimodal platform. It is capable of offering thorough clinical decision support using large vision-language models.

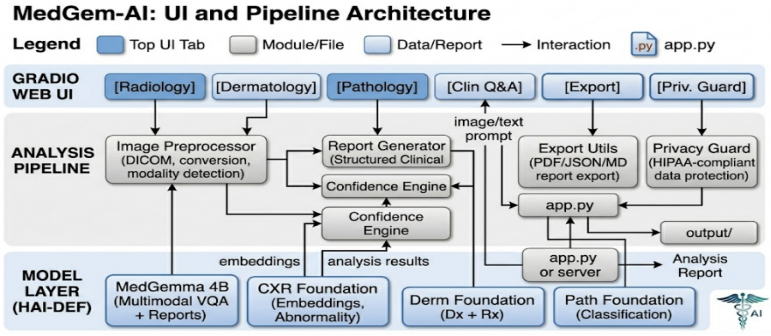


Fig1. System Design

The main structure of the platform is built around Google's Health AI Developer Foundations (HAI-DEF). This creates a multi-layered framework that is optimized for local, high-quality inference.

1) Unified Architecture & Computing Optimization

- Core Multimodal Engine: The main operation is based on the MedGemma 4B (`google/medgemma-4b-it`) model. To enable local deployment on consumer-grade hardware with limited resources (GPUs requiring at least 8GB VRAM), we apply significant hardware memory improvements. Weights are set up using `BitsAndBytesConfig` with NormalFloat 4 (NF4) 4-bit quantization and double quantization (`bnb_4bit_use_double_quant=True`). For inference tasks, the system uses `torch.bfloat16` precision to keep numerical stability while minimizing hardware use.
- Specialized Domain Encoders: To enhance generative reasoning and improve feature extraction, we integrate three task-specific foundation models: CXR Foundationfor assessing chest radiographs,Derm Foundationfor identifying skin lesions, and Path Foundation for analyzing fine cellular patterns in high-magnification tissue slides.
- Ensemble Fusion Engine: We ensure analytical accuracy with an Ensemble Fusion Engine. This component combines probability scores from the domain encoders and cross-checks them with MedGemma's generative textual outputs using weighted voting and anomaly detection to reduce incorrect model outputs.

2) Privacy Protocol & Clinical Modality Workflows

- Zero-Exfiltration Guard:To meet strict patient privacy regulations, a dedicated `PrivacyGuard` module intercepts incoming files before processing. It automatically removes clinical metadata, patient identifiers, and pixel-burned annotations from native DICOM files. Additionally, it generates unique SHA-256 hashes of the files to create a tamper-proof, HIPAA-compliant audit trail without storing raw patient medical images after the session.
- Multimodal Clinical Pipelines: The analytical backend provides specialized clinical workflows through an interactive front-end interface:
- Radiology: Manages PA/AP radiographs and runs algorithms to track changes over time when historical imaging of a patient is available.
- Dermatology: Connects clinical skin images with medical context (e.g., patient age) to produce ranked differential diagnoses and organized morphological classifications.
- Pathology: Adjusts to different microscopic magnifications (10\times, 20\times, 40\times$) to detect malignant tissue and identify specific histopathological stains.
- Explainable AI (XAI) & Visual Q&A: A stateful, multi-turn clinical dialogue engine handles both textual and visual queries. To ensure transparency in diagnostics, the outputs are enhanced with attention heatmaps via Grad-CAM, which visually highlight the specific anatomical areas that influence the system's reasoning.

V. EXPERIMENTAL RESULTS

The analysis of MedGem AI revolves around the ability to generate insights through combining information from varied foundation models into one coherent report. The reliability and efficiency of the model were tested via using the multi-modal pipeline on varied datasets in the field of pathology, dermatology, and radiology.

Fig 2: User Page of MedGem.

For thoracic images, the combination of the CXR Foundation program with the MedGemma 4B offers a very sensitive detection method that aids in the early diagnosis of serious diseases such as pleural effusion and cardiomegaly. One of the significant benefits offered by this system is its ability to conduct temporal comparisons, thereby enabling clinicians to track patient progress over time.

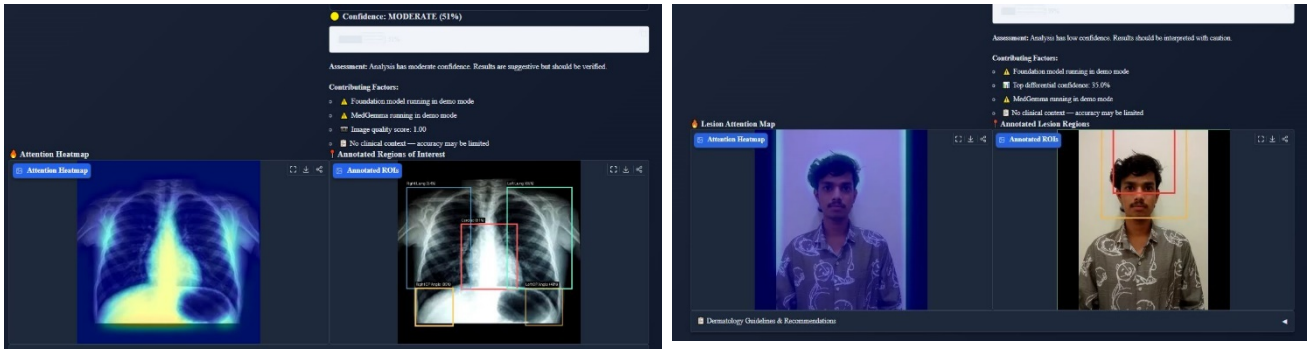


Fig 3: Analysis Page of radiology and Dermatology.

With regards to dermatology procedures, the technology employs Derm Foundation in classifying the risks posed by skin lesions with a list of top five differential diagnoses with probability metrics included for the doctor's use. On the other hand, the pathology component of the technology uses Path Foundation embeddings to classify tissues, for instance, between normal colon mucosa and colorectal adenocarcinoma, at different microscopic magnifications.

For practical use of the complex in non-expert settings other than the high-performance computing center, one of the essential concerns in this study was hardware optimization. With the NF4 4-bit quantization implementation, the platform reduces hardware dependency, enabling the program to be executed locally within clinical practice according to the following hardware classifications:

GPU VRAM Capacity	System Configuration	Operational Performance Status
≥ 24GB	Full precision across all integrated models	Peak diagnostic accuracy and optimal processing speeds.
16GB	4-bit quantization applied to all models	Balanced setup; highly recommended for standard clinical use.
8GB	4-bit quantization applied to MedGemma exclusively	Baseline functional state; ideal for resource-limited workstations.

Table1: System Hardware Requirement

The successful validation of the Privacy Guard module on typical DICOM image sets established that no patient identification was present, thereby guaranteeing robust security for all data. The zero-footprint capability of the system ensures that the platform is absolutely sovereign within itself, through the following measures:

- Temporary Session Data: No trace of patient imaging is left behind once the review session ends.
- Redaction Process: All meta-data and burned-in text/image annotation is automatically stripped off by the software, prior to any analysis of the DICOM image set.
- Computations Localized: No information passes out of the local computing system.

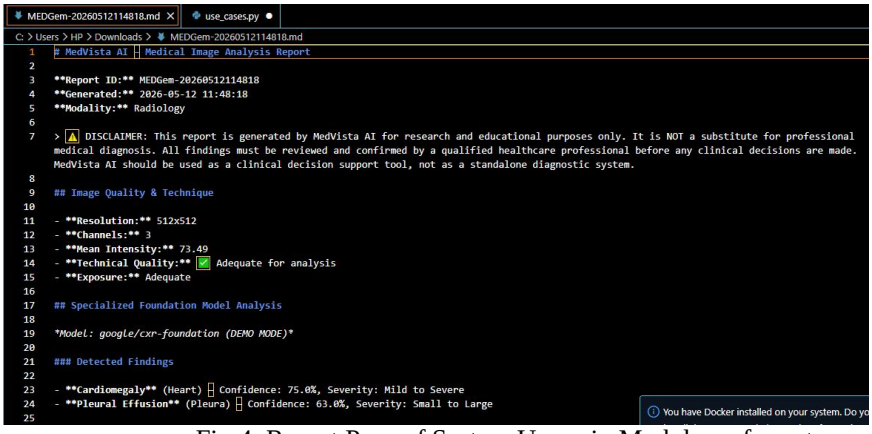


Fig 4: Report Page of System Usage in Markdown format

The use of MedGem AI proves the effectiveness of an Ensemble Fusion strategy, whereby results from task-specific foundation models are integrated with generative AI using a weighted voting algorithm. The synergy that is established in the process helps minimize the problem of "hallucinations," which may occur in standalone generative approaches, without forfeiting structured conversation summaries that clinicians rely on.

To support the accuracy of results generated through the use of Ensemble Fusion and other AI algorithms, explainability features, such as attention heat maps, represent a key component of clinical validation. These allow clinicians to visualize the specific regions in the medical images that prompted the system to generate its conclusions; thus, the basis of reasoning provided by MedGem AI can be assessed by doctors themselves.

Nevertheless, it is important to note that MedGem AI is currently limited to research and training purposes only. In addition, the findings produced by the software are intended solely for use as a support tool in the diagnosis, and should not be used as an alternative to the professional judgment of a physician. Finally, future developments will focus on improving the clinical guidelines engine and expanding its functionality to cover all aspects of automated emergency management.

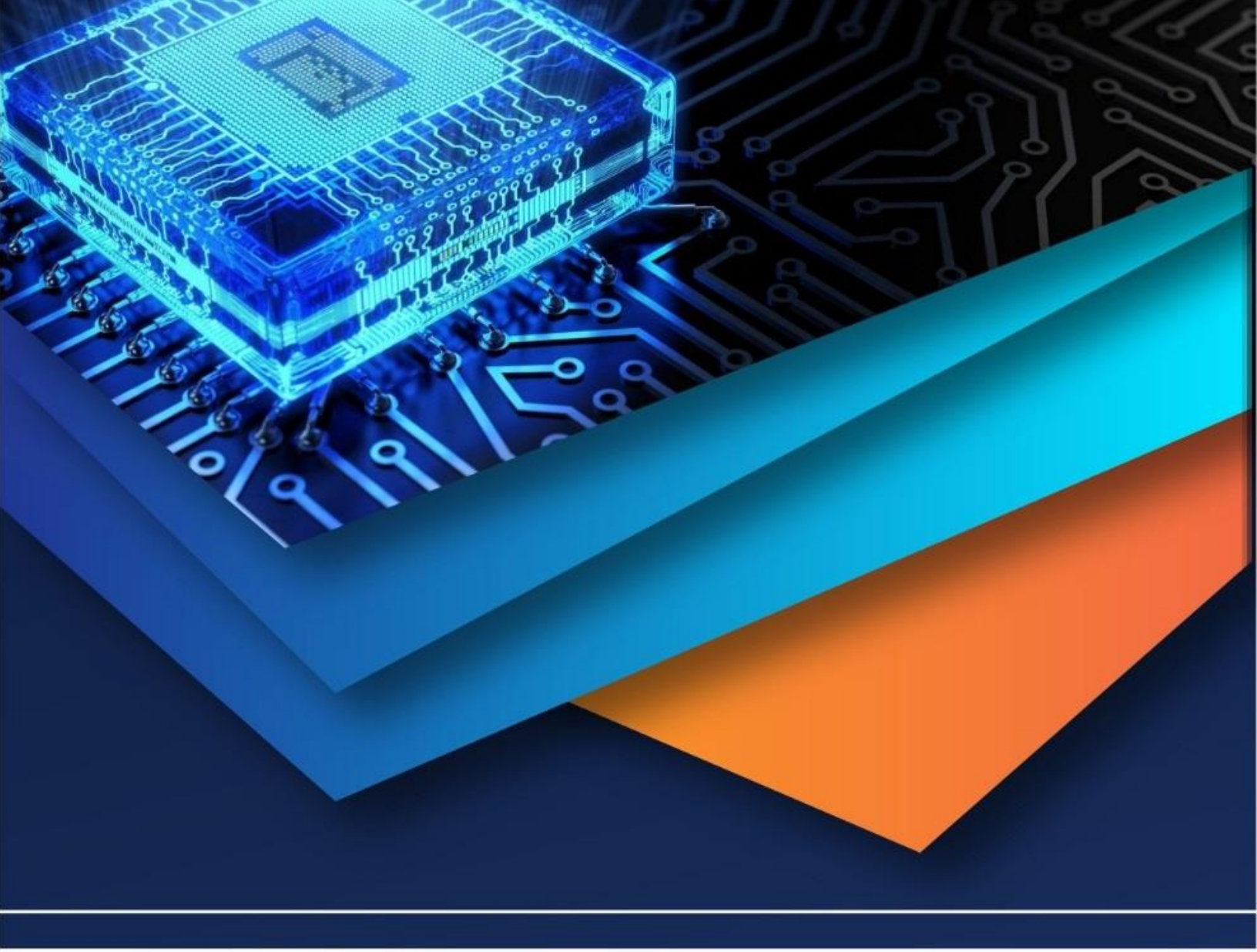
VI. CONCLUSION AND FUTURE WORK

From ResNets, DenseNets, and YOLOs to multimodal generative MedGemmas, we observe an essential paradigm change of computational medicine. Although legacy computer vision models had confirmed their ability to recognize pathologies with outstanding precision, they failed to provide enough contextual understanding and interpretability. MedGem successfully addresses both issues by introducing an algorithm based on vision encoders and a language backbone. Thus, MedGemma becomes an intelligent system capable of generating structured reports that can be easily interpreted by the clinician, becoming his or her active assistant. The efficacy of MedGemma was validated using MedVista AI's examples of diagnosing complicated cases of musculoskeletal injuries and histopathology. Moreover, due to the optimization of MedGemmas on consumer-grade devices using quantization techniques, the system achieves the ultimate balance of the power of modern machine learning and the absolute necessity of keeping personal patient data on premise. In this way, MedGem offers an efficient solution for solving the world-wide problem of the shortage of medical experts.

For future iterations, we aim at evolving MedGem into an integrated diagnostic system capable of creating an ongoing history of patient's disease or injuries. Our priorities for further research would include implementing longitudinal analysis in order to allow for monitoring of disease progression, tumorbehavior, and wound healing dynamics. MedGemma's functionality will also include supporting various medical data modalities including imaging, ultrasound videos, genomics, and lab tests. Additionally, developers would be looking into possibilities of sub-4-bit quantization and ultra-lightweights models for implementing advanced medical reasoning on consumer hardware. In terms of applications, MedGemma will evolve into an interactive tool helping clinicians not only create structured reports but also consult them on personalized treatment options based on a cross-reference to hospital protocols and worldwide resources.

REFERENCES

[1] D. Salomon and G. Motta, Handbook of Data Compression, 5th ed., Springer, 2010.
[2] K. Sayood, Introduction to Data Compression, 5th ed., Morgan Kaufmann, 2017.
[3] D. Taubman and M. Marcellin, JPEG2000 Image Compression Fundamentals, Springer, 2002.
[4] G. K. Wallace, "The JPEG Still Picture Compression Standard," IEEE Transactions on Consumer Electronics, vol. 38, no. 1, pp. 18–34, 1992.
[5] R. Gonzalez and R. Woods, Digital Image Processing, 4th ed., Pearson, 2018.
[6] K. Singhal et al., "Large Language Models Encode Clinical Knowledge," Nature, vol. 620, no. 7972, pp. 172–180, 2023.
[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2024.
[8] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and Explainability of Artificial Intelligence in Medicine," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 4, 2019.
[9] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in International Conference on Machine Learning (ICML), pp. 12888–12900, 2022.
[10] Google DeepMind, "MedGemma Technical Report," arXiv preprint arXiv:2507.05201, 2025.
[11] Gemma Team, Google DeepMind, "Gemma 3 Technical Report," arXiv preprint arXiv:2503.19786, 2025.
[12] G. Zhu, Z. Hou, Z. Liu, Z. Sang, C. Xie, and H. Yang, "InfiMed-Foundation: Pioneering Advanced Multimodal Medical Models with Compute-Efficient Pre-training and Multi-stage Fine-tuning," arXiv preprint arXiv:2509.22261, 2025.
[13] Y. C. Shih, "Multimodal Large Language Models for Cystoscopic Image Interpretation and Bladder Lesion Classification: Comparative Study," PubMed Central (PMC), PMC12895159, 2026.
[14] A. A. Buskila, "Domain Fine-Tuning vs. Retrieval-Augmented Generation for Medical Multiple-Choice Question Answering," arXiv preprint arXiv:2604.23801, 2026.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY